DOCUMENT RESUME

ED 191 904 TM 800 528

AUTHOR Shaver, James P.
TITLE Readdressing the Role of Statistical Tests of Significance.
PUB DATE Apr 80
NOTE 23p.: Paper presented at the Annual Meeting of the American Educational Research Association (64th, Boston, MA, April 7-11, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Data Analysis: *Educational Research: Graduate Study: Higher Education: *Sampling: *Statistical Bias: *Statistical Significance: Validity
IDENTIFIERS *Inference (Statistical)

ABSTRACT

The implications of data from a review of ten years of the American Educational Research Journal (AERJ) indicating that random sampling is rare and that there is increased use of quasi-experimental designs lacking in random assignment are considered. It is suggested that tests of significance could be abandoned or at least placed in a subsidiary role as a test of one rival internal validity hypothesis (i.e., that chance caused the results) rather than being the center of data analysis. The use of a priori effect size criteria for evaluating the importance of research results, the greater use of measures of association, and more replications are advocated. A shift in emphasis from teaching graduate students in education how to do tests of significance to the examination of the theory behind inferential statistics could have a salutory effect on educational research. (RL)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. Shaver
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED 191904

TM 800528

READDRESSING THE ROLE OF STATISTICAL TESTS OF SIGNIFICANCE*

James P. Shaver
Utah State University

Educational researchers are often criticized as atheoretical and ahistorical (e.g., Clifford, 1973)--that is, for not using their research to test theory and for not basing their studies on past research. Although the criticisms are typically leveled at the choice and formulation of research problems and objectives, they could just as well be aimed at the selection and use of methodology, especially the use of inferential statistics.

The intent, when this symposium was first proposed, was that this paper would address the question, "What can tests of statistical significance tell us?" The answer was to be, "Not nearly as much as has typically been assumed by educational researchers." That answer was to be elaborated on with a sketching of the randomness assumptions in the statistical theory underlying the inferential statistics commonly used in educational research, some commentary on the restricted nature of the questions answered by inferential tests, a mentioning of some data that indicate that, in any event, the assumptions are rarely met, and some suggestions for alternatives to inferential statistics. That general intent will still be followed.

In preparing to write this paper, however, I was very much impressed by a quite different point on which I would like to comment in passing: That is, educational research would be a fruitful field of study for those interested in the history and sociology of ideas. To try to chart and explain the rooting, growth, and unexamined acceptance and use of inferential statistics which, at best, do not answer the questions

*Paper prepared for a symposium, Tests of Significance: Readdressing their Role, annual meeting of the American Educational Research Association, Boston, April 7, 1980.

to which they are frequently put, and which may well have contributed thereby to thousands of relatively meaningless research studies and reports would be a fascinating study. How did statistical tests of significance become such an ingrained element in the educational research culture when they seem to be so insignificant (conceptually, not statistically) a part of the natural and biological sciences (Shaver, 1979, p. 5) which those in the social sciences and their stepsister, educational research, have tried, even if not wholeheartedly, to emulate? Why have the criticisms of tests of statistical significance in the psychological and sociological literature of the 1950's and 60's (see, e.g., the papers in Morrison & Henkel, 1970a, as well as Skinner, 1956; Bakan, 1967; Seeman, 1973; Signorelli, 1974) had such slight effect on educational (as well as psychological and sociological) research? How many graduate programs expose students to such readings (as well as more recent ones, such as Carver, 1978) or include courses that critically examine the assumptions and limitations of inferential statistics? How many of those now teaching research and statistics courses and advising graduate students come from such programs? How many supervisory committees will not accept proposals for thesis or dissertation research unless traditional data-gathering and tests of statistical significance are included? Surely, the study of inferential statistics in educational research would be a fruitful area for those interested in intellectual history and the sociology of ideas. But back to the main theme.

Teaching Inferential Statistics

An appropriate context for readdressing the role of statistical tests of significance in educational research is the statistics course(s)

taught to graduate students by many AERA members, including me. Examining how a course (or course segment) in inferential statistics might be taught will bring into focus some issues about the use of inferential statistics and some suggestions for alternative research strategies.

Should Inferential Statistics be Dropped?

Given the severe criticisms of statistical tests of significance by individuals such as Carver (1978), one might ask if we should continue teaching inferential statistics to our graduate students. Of course, we must continue to do so. Most graduate faculty--even, perhaps especially, those with the least knowledge and experience in research and statistics--would not even entertain discontinuance. The use of inferential statistics is such a strong element in the culture of educational research that they fear, as Bakan (1967, p. 13) has noted, the entire tapestry will come apart if that one thread is unraveled. They may be right. (And perhaps the tapestry needs to be totally unraveled and reconstructed with some new threads.) Moreover, graduate students not schooled in inferential statistics would not be able to read and critique the majority of educational research reports to-date nor speak the langugage of educational research that permeates the discussion of educational issues generally. Despite whatever alternate research conceptualization strengths they might have, such graduate students would likely have a difficult time finding employment. To so disadvantage them would not only be unfair, but would probably threaten the existence of any one graduate program which was so innovative on its own (except for those with existing strong specialization

in a nonstatistically-oriented methodology, such as field studies or ethnography). Moreover, statistical tests do have some utility, although not of the extensive nature often assumed or implied. What to do with the inferential statistics course, then?

The Legitimate Uses of Inferential Statistics

Graduate students, and other researchers, often assume that inferential statistics will tell them too much: Whether a null hypothesis is true in a population, whether a treatment is effective, whether a finding would re-occur if a study were replicated or if the treatment which was "investigated" was used generally in the field. Such misconceptions, because they are prevalent and dysfunctional, should be central concerns of instructors of statistics courses.

Theoretical Distributions and Randomness. Two basic uses for inferential statistics are often advocated--assisting the researcher in making decisions about (a) the generalizability of findings (external validity) and (b) the effectiveness of treatments (internal validity; Campbell & Stanley, 1963). For example, Borg and Gall (1979, pp. 421-2) discussed the role of statistical inference in deciding whether sample findings are applicable to a population. And, Winch and Campbell (1969) maintained that the test of significance provides "a relevant and useful way of assessing the relative likelihood that a real difference exists and is worthy of interpretive attention . . ." (p. 140). Students need to confront the assumptions underlying both uses, as well as the rather severe restrictions.

The theoretical sampling distributions typically used in inferential statistics in educational research (normal, *F*, *t*, Chi-square) are based

on the assumption of randomness--that the distributions would be approximated if a large number of random samples were drawn. Random sampling (which is apparently not common in educational research; see Shaver & Norton, 1980a, 1980b) "insures constant and independent probabilities" (Winer, 1971, p. 5), making it possible to assume that each sample "is randomly representative of the population in all respects" (Glass & Stanley, 1970, p. 242, italics theirs). If one hopes to reason back to a population, random selection from a clearly defined population is essential (see Morrison & Henkel, 1970b). Bracht and Glass (1968), Kerlinger (1975, p. 206), and Bock (1975, p. 236) are among those who have noted the lack of adequate basis for estimating population parameters (either point or interval estimation) when samples are not random.

The use of inferential statistics to generalize to populations without random sampling is, then, questionable, although a common practice in educational research. The effects of violating the assumptions of normality and homoscedasticity which underlie our use of the common parametric *t* and *F* tests of significance could be examined once high speed computers were available (and the conclusion could be drawn that the tests are robust, with levels of significance affected greatly only when variances are not homogeneous and sample sizes are unequal--see, e.g., Glass & Stanley, 1970, pp. 388-74). However, the effects of violating the assumption of randomness (sampling or assignment)--or even the extent of the violations--cannot be so easily examined. Unlike the assumptions of normality and homoscedasticity which have to do specifically with the dependent variable(s), there is no such closed

set of variables to be presumed of concern for studying the effects of nonrandomness. The importance of randomness lies in the distribution of all present variables according to chance. To enumerate every potentially relevant variable and specify its relationship to the dependent variable(s), being certain no crucial variable was overlooked, in order to investigate the effects of nonrandomness on probability statements presents insuperable difficulties.

Despite occasional misleading statements in statistics textbooks (e.g., Winer, 1971, p. 752[1]; Ferguson, 1976, p. 346), random sampling to groups is an assumption underlying the sampling distributions not only for the t-test and analysis of variance, but for analysis of covariance (Elashoff, 1969; Glass & Stanley, 1970, p. 340) and nonparametric statistics (Siegel, 1956, e.g., pp. 32, 48, 78, 88, 152; Morrison & Henkel, 1970b). When random assignment to groups occurs without random sampling, the groups cannot be assumed to be randomly representativeness of any population of interest. And, when random sampling is not done, a serious question must be raised about the validity of the assumption that the sample variance is the best estimate of the population variance as used to calculate the standard error of mean differences or the error estimate. The possibility of a systematic bias, as contrasted with a chance deviation from the population value, suggests that the meaning of the resultant probability statement may be doubtful.

Random assignment does insure pre-experimental equivalence, in the sense of eliminating all but chance differences on all present variables

---

[1]Interestingly, in his examples of applications of analysis of covariance, Winer uses random sampling and assignment (pp. 776, 808).

(Kerlinger & Pedhauser, 1973, p. 82; Gilbert, Light & Mosteller, 1975, pp. 136-7). For that reason, random assignment with or without random sampling is an important method of controlling several threats to internal validity (Campbell & Stanley, 1963), especially maturation, regression, and selection. Campbell and Boruch (1975) have discussed at length the potential difficulties in interpretation arising from lack of random assignment, and lamented (p. 202) the increased use of quasi-experimental designs, lacking in randomization.

Whether random assignment is essential to the actual use of tests of significance, a question usually avoided in statistics textbooks, is a matter of some dispute in the literature. For example, Morrison and Henkel (1970b, pp. 305-6) emphasize the importance of meeting the randomness assumption. Others, such as Winch & Campbell (1969), argue that inferential tests provide excellent approximations of the results that are obtained when exact probabilities are calculated based on random assignments of the actual obtained data. It is argued that inferential statistics, such as the t-test, can therefore be used to answer the question, how likely is it that the obtained difference would have occurred by chance if random assignment to groups had occurred? (See Winch & Campbell, 1969.)

Unfortunately, the effects of nonrandom sampling on calculated test of significance probabilities are not well known. Berk and Brewer (1978) argue that such techniques confound any treatment effect in the comparison, so that in cases where there is a difference, too many decisions will be made that it is likely due to chance. They suggest that the exact randomized component model is appropriately used with research populations, but not for the usual test of sample differences.

In any event, exact tests (such as Fisher's randomization tests of location, Bradley, 1968), which are distribution-free in the sense that they involve generating a probability statement by considering all permutations of observations for the particular sized groups used in the research, are of no use in generalization if the groups are not random samples. And, such tests are little known and little used in educational research, probably in large part because of the computational difficulties. Moreover, if subjects have not been assigned randomly to groups, one loses the basic assumption of pretreatment equivalence.

For the reasons noted above, without randomization probability statements based on our common inferential statistics are indeterminate.[2] That inadequacy should be strongly emphasized in statistics courses. The purpose should be to discourage the common practice of not drawing random samples or randomly assigning subjects (especially in light of the randomization possibilities in the schools--see, Campbell and Boruch, 1975) without some research strategy in mind other than one-study data collection and analysis with inferential statistics. An equally important purpose should be to lead graduate students to the consideration of methods other than inferential statistics with which to draw conclusions from data.

Limitations Even with Randomness. It may seem too obvious to mention that even with true experimental studies involving randomization, evidence, based on inferential statistics, that a finding of a difference between means favoring the treatment is significant is not proof that the treatment was effective. Yet, such claims are not uncommon, particularly in theses and dissertations. Aside from the rather apparent point that one may be

---

[2]Other sources of indeterminacy, such as the lack of independent observations as units of analysis, should not be ignored either.

making a Type I error in drawing the conclusion that the finding--the difference between means, frequencies, or whatever--was not due to chance, obviously the statistical test only speaks to chance as an alternative explanation. Potentially, several sources of lack of internal validity may have produced the result, and that should at least be alluded to in a statistics class.

Typically, it is difficult for readers of research reports to determine how threats to internal validity, other than chance, were handled. Carver (1978) has discussed the tendency for educational researchers to focus exclusively on chance as an alternative explanation, rather than treating it as one alternative that needs to be checked out along with others. At least in part because of that focus on the dependent variable(s) and tests of significance, design and procedures tend not to be carefully analyzed for possible effects on outcomes nor carefully described so that readers can reason about what conclusions are warranted. (The lack of description is also, of course, an impediment to replication--or would be if educational researchers were inclined to replicate, which they seem not to be, Shaver & Norton, 1980a, b; Cronbach & Snow, 1977, p. 16.) Unfortunately, designing, conducting, and reporting research seem to get little treatment in statistics classes, and the link (or lack of link) between statistical significance and causal hypotheses also does not seem to be a point accented in design classes. One point of emphasis in teaching inferential statistics should be the limited role of the statistical test of significance, and the importance of design considerations, in drawing cause-and-effect conclusions.

What if having set an Alpha level, usually by accepting the traditional .05 (a topic that deserves considerable attention in statistics classes), and having conducted an inferential test on random samples, the researcher is able to conclude that a difference in means is not likely due to chance? What does this tell him/her about the population? Perhaps nothing, because it may be an erroneous decision (Type I error). Too often, inferences are drawn on one piece of evidence of statistical significance. And, as Berk and Brewer (1978) have pointed out, tests of significance deal only with what can be expected in the long run and do not tell one whether a particular result is due to chance or not.

Tukey (1969) has noted that our modern use of the statistical test of significance is largely due to R.A. Fisher. But, Tukey noted:

> Sir Ronald's standard of firm knowledge was not one very extremely significant result, but rather the ability to repeatedly get results significant at 5% (p. 85).

"Variability and significance and confidence," Tukey (1969, p. 85) pointed out, can be judged only on the basis of repetition. The need

for replication must be emphasized in statistics courses as well as design courses if educational research is ever to move from the one-shot type of research strategy (or lack of strategy) that seems to be dominant.

The researcher's statistically significant finding must also be considered in light of the tendency to state the result in terms of "rejecting the null hypothesis". In the first place, to "reject the null hypothesis" makes no sense. In making a test of statistical significance the researcher uses a theoretical distribution based on the assumption of random sampling from a population (or populations) in which the null hypothesis is true (for example, a normal distribution of mean differences based on repeated drawing of large pairs of random samples in which the mean difference between means is zero). While it is legitimate to conclude (if the researcher has random samples, or at least has used random assignment) that a finding is an unlikely occurrence, to reject the null hypothesis is nonsensical, because it is only an assumption which allows a statement of probability in relation to a sampling distribution.

Experienced researchers recognize that illogic, I am sure; but graduate students who are allowed to "reject the null hypothesis" often are lead, I fear, to the erroneous conclusion that they have thereby said something about the population from which their samples came. The inferential test tells us how likely it is that the sample come from a population in which the null hypothesis was true, but does not, conversely, tell us anything further about the population (see Berkson, 1970; Morrison & Henkel, 1970b). As Carver (1978) has pointed

out in his discussion of the "odds-against-chance fantasy," the inferential test does not tell the researcher that the result is not due to chance; it tells only the probability of the result occurring if the null hypothesis is assumed to be true. Once the researcher concludes that it is not likely that his/her results occurred under the null hypothesis (i.e., that the theoretical sampling distribution does not apply), he/she loses the basis for stating how likely it is chance caused the effect. That would require a new, unavailable sampling distribution based on the hypothesized population value. Carver (1978, pp. 384-85) likens the error of drawing the conclusion that a statistically significant result tells us the probability that two samples did not in fact come from the same population to substituting our rather certain knowledge that a hanged person is dead for an estimate of the probability that a dead person was hanged. (Also see Berkson, 1970.) Our statistical test tells us the probability of our finding if the null hypothesis is assumed to be true, not the likelihood that the null hypothesis is true given our finding.

Confidence Intervals

Some advance in conceptual clarity might be made if we would urge graduate students to use confidence intervals rather than traditional tests of significance (Cronbach & Snow, 1977, pp. 53, 86). That is, sample data (assuming random sampling) could be used to set up a range within which (at some defined and, hopefully, defended level of confidence) it is predicted that the population value falls. Although the "odds-against-chance fantasy" is not automatically avoided by setting confidence intervals, it at least moves the researcher to

thinking about a range of possibilities, some of which may be very close to the null hypothesis even though it is not encompassed in the interval. And, replication of findings to determine if they fall in that interval would still be crucial for establishing confidence in one's results. Although estimating confidence intervals and tests of significance are two sides of the same coin (Glass & Stanley, 1970, p. 268), more emphasis on interval estimation might move graduate students away from the dichotomous "yes-no" thinking that is typical when tests of significance are used as a basis for decision-making.

Other Criteria

Statistics courses also ought to deal extensively with the notion of judging the significance of research results by other preset criteria than an Alpha level. If graduate courses urged researchers to consider, prior to carrying out a piece of research, how large an effect would be practically, educationally, or theoretically[3] important, and then to use those prior criteria as a basis for judging outcomes, not only might statistical significance lose some of its allure but an important dimension would be added to decision-making about the significance of findings.

Making decisions about the size an effect must reach to be important is usually not easy in educational research, for a number of reasons. One is that current research paradigms have not conditioned us to think that way. We are conditioned to set (accept) an Alpha level and assume

---

[3]Moving from theory to predicted result, rather than trying to use statistics to determine "what reality is like" and generate theory, would make educational researchers much more allied with natural scientists, who use mathematics to test and add preciseness to theoretical formulations. (See Easley, 1973, p. 56; Signorelli, 1974; Freedman & van Nieuwenhuizen, 1978.)

statistical significance is of overwhelming importance, with practical or educational significance a secondary matter. (Theoretical significance is rarely even considered.) Also, such decisions often involve value judgments, and we have not developed intellectual strategies for identifying and weighing values as part of our research paradigm (see Larkins & Shaver, 1972). Related to the above point is the fact that educational objectives are rarely stated with such clarity and precision (and perhaps often cannot and should not be) as to provide a solid basis for formulating acceptable criteria of significance. Moreover, the tests we use as dependent variables often lack not only a value-base, but are so imprecise in meaning that interpretation in regard to importance is very difficult. All of these considerations are, of course, interrelated. How does one decide how many points change on a mathematics test is important? What will five points of change mean in terms of valued mathematics knowledge or performance and/or, for example, future school or professional/vocational performance? Is that change worth the cost of the treatment which is to effect it? Such questions are even harder to answer with a variable such as self-concept.

Power Analysis

What if the researcher's result reached the preset educational importance criterion, but was not statistically significant? Did she/he simply err in not working with a large enough sample to detect the anticipated effect--an answer suggested by the power analysis approach to statistical analysis (Cohen, 1977).

Power analysis has not been widely used by educational researchers (Cohen & Hyman, 1979), perhaps in part because of the difficulty of pre-specifying desired effects. To use power analysis, one must deal with effect size, Alpha, Beta (to arrive at the chance of detecting a difference if it exists, $1-\beta$--that is, the desired power), and N. Not only does power analysis, then, call for pre-data gathering judgments that researchers are not used to making because such decisions are not part of the prevalent paradigm, but it still relies on statistical significance, as Cohen and Hyman (1979, p. 15) make clear. They note that if, for example, adopting an Alpha level other than .05 to increase power seems deviant, the researcher can juggle N to get desired power.

While more discussion of power analysis in inferential statistics courses might make graduate students more conscious of elements in inferential statistical decision-making which tend to get ignored, it will not avoid the problem of meaningless statistical results. Moreover, specifying an effect size, Alpha, and Beta, in order to discover the N one must have to find a significant effect seems more akin to gaming than to building sound research knowledge. However, that approach does make it clear that statistical significance is a function of sample size. And, if one is simply willing to obtain large enough samples, almost any effect (assuming reliable measurement) can be found to be significant (see, e.g., Meehl, 1970, p. 259).

Inference and Association

In recommending that graduate students be encouraged to think of statistical tests of significance as only one criterion against which

to judge results, and that pre-specified effects also be used as standards, I have referred to effects as anticipated differences. Of course, there are other effects of interest, such as correlations between variables. And difference effects can be expressed as other than differences between means--e.g., as the percentage of increase in one group's performance as compared to another group's or that the difference will equal or exceed the standard deviation for the dependent variable in the control group. Another useful approach for estimating effect, or for adding information for the post hoc interpretation of findings, is to think in terms of the amount of variance in the dependent variable that is associated with the independent variable.

The interpretation of r^2 (or R^2) in terms of proportion of explained or common variance is a common topic in statistics courses, as is, hopefully, the idea that when N's are large, statistically significant r's may not indicate much common variance.[4] That same line of reasoning is, of course, applicable to the testing of significant mean differences with *t* or *F* tests. Even with a significant *F*, the treatment may not have accounted for any but a trivial portion of the variance on the dependent variable. The importance of using a measure of association such as Eta^2 (the correlation ratio) to obtain an estimate of variance accounted for by the treatment should be emphasized in statistics courses (see, e.g., Hayes, 1973, Ch. 12; Ferguson, 1976, pp. 236-7, 473).

---

[4] Researchers still occasionally report significance levels for r's without reporting the r's themselves, which suggests that sometimes the idea is not taught or understood. Of course, interpreting statistically significant correlation coefficients calls for attention to the same problems as have been discussed earlier in terms of mean differences.

Of course, Eta^2 is open to misinterpretation, too. The variance accounted for may be due to some extraneous variable, not the treatment. Or, interpretation may not be appropriately limited to the particular categories of the independent variable used in the study (see, e.g., Glass & Hakstian 1969, pp. 409-10). But knowledge of such measures of association will provide graduate students with another set of powerful intellectual tools, to be used either in pre-specifying effects or in the post hoc interpretation of findings--e.g., when a large N makes one wonder if a finding is statistically significant but patently trivial.

Data Exploration

Another helpful addition to statistics courses would be to encourage students to look at their data, rather than (and certainly prior to) simply running statistical tests of significance on descriptive statistics--the almost exclusive focus of attention now. Tukey (1977) is among those advocating this revolutionary idea. The use of the mean, because the available tests of significance make it a convenient summary statistic, probably distorts impressions of data more than most of us realize. Attention to and reporting of other measures of central tendency, of estimates of variability, even of frequency distributions, might be the result of data exploration. Another offshoot would be increased recognition of the frequently inappropriate application of our common inferential statistics to specific sets of data and the invalidity of many of the conclusions drawn through inferential statistics applied to means (and even some of the conclusions that might be drawn using the usual descriptive statistics to evaluate educational significance).

Conclusion

The use of statistical tests of significance has been too long accepted as an unexamined ritual in educational research. While such tests can play a legitimate role in dealing with the alternative hypothesis of chance, they have been overused and misused. The limited information provided the researcher by inferential statistics, even when the fundamental assumption of randomness is met, is not often recognized in research reports. Even less often is there acknowledgement that the failure to sample and/or assign randomly makes probability statements indeterminate.

The preparation of this paper did not include a survey of the content of statistics courses for graduate students in education. However, based on personal experience as a student and instructor, on comments during conversations and correspondence with colleagues, and on observations while reading the research literature, it appears that much more could be done in graduate statistics courses to engage future educational researchers in readdressing the role of inferential statistics. A shift in emphasis from teaching students how to do tests of significance to the examination of the theory behind inferential statistics, especially the central place of randomness and the limited conclusions to be drawn from tests of significance, accompanied by the consideration of alternatives for making educational (or practical or theoretical) significance an explicit standard for judging findings, could have a salutory effect on educational research.

The use of inferential statistics has too often, I believe, become a substitute for carefully defining research problems, thinking about design, and considering outcomes. Statistics courses ought to help students put inferential statistics in perspective as one, limited research tool--a tool which cannot answer cause-and-effect questions, which cannot tell one much about populations, and nothing about them without random sampling, which does not preclude the importance of repetition of findings, and which is no substitute for thoughtfulness about the types of findings that will have theoretical or educational significance. The history and sociology of ideas to explain the paramount position of inferential statistics in educational research, despite its limited usefulness, has not yet been written. But if it is ever done, graduate program instruction will, I believe, be found to be a major, although not sole, explanation.

Such instruction could also be cited in the future as a major factor in a shift away from over-reliance on statistical inference in research, with all its attendant shortcomings.[5] If that citation is to occur, however, the hold of the statistical Zeitgeist must be broken. Breaking loose will not be an easy task because of our own backgrounds and the view that continues to dominate textbooks, journals, and many of our colleagues. The hope behind this paper is to encourage the necessary examination of assumptions--a scrutiny that appears to be long overdue.

---

[5] Rist (1980), in discussing the recent growth in ethnographic-type studies in education, has noted the growing skepticism about quantitative methods among educational researchers. He also cautions that as other methods emerge, they, too, may become dysfunctionally dominant so that, as with the quantitative paradigm, problems will be selected and shaped to method rather than the selection of method being based on appropriateness for defined problems.

References

Bakan, D. On method: Toward a reconstruction of psychological investigation. San Francisco: Jossey-Bass, 1967.

Berkson, J. Tests of significance considered as evidence. In D.E. Morrison and R.E. Henkel (Eds.), The significance test controversy--A reader. Chicago: Adeline, 1970.

Bock, R.D. Multivariate statistic methods in behavioral research. New York: McGraw-Hill, 1975.

Borg, W.R., & Gall, M.D. Educational research: An introduction (3rd ed.). New York: Longman, 1979.

Bracht, G.F., & Glass, G. V. The external validity of experiments. American Educational Research Journal, 1968, 5, 437-474.

Bradley, J.V. Distribution-free statistical tests. Englewood Cliffs, N.J.: Prentice-Hall, 1968.

Campbell, D.T., & Boruch, P.F. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C.A. Bennett & A.A. Lumsdaine (Eds.), Evaluation and experiments: Some critical problems in assessing social programs. New York: Academic Press, 1975.

Campbell, D.T., & Stanley, J.C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Carver, R.P. The case against statistical significance testing. Harvard Educational Review, 1978, 48, 378-399.

Clifford, G.J. A history of the impact of research on teaching. In R.M.W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally, 1973.

Cohen, J. Statistical power analysis for the behavioral sciences. (Rev. Ed.). New York: Academic Press, 1977.

Cohen, S.A., & Hyman, J.S. How come so many hypotheses in educational research are supported? (A modest proposal). Educational Researcher, 1979, 8(11), 6-11.

Cronbach, L.J., & Snow, R.E. Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington, 1977.

Easley, J.A., Jr. The natural sciences and educational research: A comparison. In H.S. Broudy, R.H. Ennis, and L.I. Krimerman (Eds.), Philosophy of educational research. New York: Wiley, 1973.

Berk, R.A., & Brewer, M. Feet of clay in hobnail boots: An assessment of statistical inference in applied research. In Thomas D. Cook (Ed.), Evaluation Studies Review Annual, 1978, 3, 190-214.

Elashoff, J.D. Analysis of covariance: A delicate instrument. American Educational Research Journal, 1969, 6, 383-401.

Ferguson, G.A. Statistical analysis in psychology & education (4th ed.). New York: McGraw-Hill, 1976.

Freedman, D.Z., & van Nieuwenhuizen, P. Supergravity and the unification of the laws of physics. Scientific American, 1978, 238, 126, 128-9, 131-9, 141-3.

Gilbert, J.P., Light, R.J., & Mosteller, F. Assessing social innovations: An empirical base for policy. In C.A. Bennett & A.A. Lumsdaine (Eds.), Evaluation and experiment: Some critical problems in assessing social programs. New York: Academic Press, 1975.

Glass, G.V, & Hakstian, R. Measures of association in comparative experiments. American Educational Research Journal, 1979, 6, 403-414.

Glass, G.V, & Stanley, J.C. Statistical methods in education and psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Hayes, W.L. Statistics for the social sciences (2nd ed.). New York: Holt, Rinehart and Winston, 1973.

Kerlinger, F.N. Foundations of behavioral research (2nd ed.). New York: Holt, Rinehart & Winston, 1973.

Kerlinger, F.N., & Pedhauser, E.J. Multiple regression in behavioral research. New York: Holt, Rinehart & Winston, 1973.

Larkins, A.G., & Shaver, J.P. "Hard-nosed" research and the evaluation of curriculum. Teachers College Record, 1972, 73, 415-422.

Meehl, P.E. Theory testing in psychology and physics: A methodological paradox. In D.E. Morrison and R.E. Henkel (Eds.), The significance test controversy--A reader. Chicago: Adeline, 1970.

Morrison, D.E., & Henkel, R.E. (Eds.). The significance test controversy--A reader. Chicago: Adeline, 1970. (a)

Morrison, D.E., & Henkel, R.E. Significance tests reconsidered. In D.E. Morrison and R.E. Henkel (Eds.), The significance test controversy--A reader. Chicago: Adeline, 1970. (b)

Rist, R.C. Blitzkrieg ethnography: On the transformation of a method into a movement. Educational Researcher, 1980, 9(2), 8-10.

Seeman, J. On supervising student research. American Psychologist, 1973, 28, 900-906.

Shaver, J.P. The productivity of educational research and the applied-basic research distinction. Educational Researcher, 1979, 8(1), 3-9.

Shaver, J.P., & Norton, R.S. Populations, samples, randomness, and replication in two social studies journals. Manuscript submitted for publication, 1980. (a)

Shaver, J.P., & Norton, R.S. Randomness and replication in ten years of the American Educational Research Journal. Educational Researcher, 1980, 9(1), 9-15. (b)

Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Signorelli, A. Statistics: Tool or master of the psychologist? American Psychologist, 1974, 29, 774-777.

Skinner, B.F. A case history in scientific method. American Psychologist, 1956, 11, 221-233.

Tukey, J.W. Analyzing data: Sanctification or detective work? American Psychologist, 1969, 24, 83-91.

Tukey, J.W. Exploratory data analysis. Reading, MA: Addison-Wesley, 1977.

Winch, R.F., & Campbell, D.T. Proof? No. Evidence? Yes. The significance of tests of significance. American Sociologist, 1969, 4, 140-143.

Winer, B.J. Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill, 1971.